COMPLETE STUDY GUIDE

Google Data Analytics Certificate

Complete Study Guide

120+ Pages	8 Chapters	100 Practice Qs	Lifetime Access

Table of Contents

Bonus Resources

- SQL Quick Reference Sheet
- Spreadsheet Formula Cheat Sheet
- Capstone Project Template
- Data Analyst Resume Guide

Chapter 1

Foundations of Data Analytics

The data lifecycle, analyst roles, and the tools of the trade

What Do Data Analysts Actually Do?

Data analysts collect, clean, and interpret data to help organizations make better decisions. The work spans industries — from healthcare to retail, finance to tech — but the core responsibilities are consistent: turning raw data into actionable insights that stakeholders can use.

Google defines six key tasks that data analysts perform: ask, prepare, process, analyze, share, and act. This framework — known as the data analysis process — runs through the entire certificate and provides a structure you'll use throughout your career.

The Data Analysis Process

The Six Phases

Ask → Prepare → Process → Analyze → Share → Act. Every analytics project follows some version of this framework, regardless of industry or tool.

Phase	What You Do
Ask	Define the problem and the questions your analysis needs to answer. Understand stakeholder needs.
Prepare	Identify and collect the data you need. Assess its credibility and bias.
Process	Clean and transform data into a usable format. Document every change.
Analyze	Use tools (SQL, spreadsheets, R) to find patterns and draw conclusions.
Share	Present findings using visualizations and clear narratives.
Act	Recommend actions based on your analysis. Help stakeholders implement decisions.

Data Analyst Roles and the Data Ecosystem

Understanding where you fit in the broader data ecosystem is important for interviews and for doing your job well. The main roles you'll encounter:

- Data Analyst — collects, cleans, and analyzes data to inform decisions. Entry-level accessible with the GDA certificate.
- Data Scientist — builds predictive models and uses machine learning. Typically requires advanced math and programming.
- Data Engineer — builds and maintains the infrastructure that stores and moves data. Focuses on pipelines, not analysis.
- Business Intelligence Analyst — specializes in dashboards, reporting, and business metrics.
- Database Administrator — manages the systems that store data. Less analysis-focused.

Key Tools Covered in the Certificate

Category	Tools	Used For
Spreadsheets	Google Sheets / Excel	Data entry, cleaning, VLOOKUP, pivot tables
SQL	BigQuery, MySQL	Querying databases, filtering, aggregating data
Visualization	Tableau Public	Dashboards, charts, data storytelling
Programming	R / RStudio	Statistical analysis, ggplot2 charts

Data Types and Structures

You'll encounter two main categories of data throughout the certificate:

Quantitative vs. Qualitative

Quantitative data is numerical and measurable — sales figures, temperatures, page views. It can be added, averaged, and statistically analyzed. **Qualitative data** describes qualities or categories — customer feedback, product names, geographic regions. It can be counted and categorized but not mathematically operated on directly.

Structured vs. Unstructured

Structured data is organized in rows and columns — like a spreadsheet or database table. It's easy to query with SQL. **Unstructured data** has no predefined format — emails, images, audio files, social

media posts. Analyzing it requires different tools and techniques not covered in the GDA certificate.

Chapter 1 Practice Questions

Question 1

Which phase of the data analysis process involves identifying and collecting the data you need?

A. Ask
B. Prepare
C. Process
D. Analyze

✓ Correct Answer: B

Prepare is the phase where you identify data sources, collect data, and evaluate its credibility and potential bias.

Question 2

What is the primary difference between a data analyst and a data scientist?

A. Data analysts use SQL; data scientists do not
B. Data analysts focus on answering business questions with existing data; data scientists build predictive models
C. Data scientists work with smaller datasets
D. There is no meaningful difference between the roles

✓ Correct Answer: B

Data analysts primarily use existing data to answer defined business questions. Data scientists build predictive and prescriptive models, often requiring advanced statistics and machine learning expertise.

Question 3

Which of these is an example of qualitative data?

A. Monthly sales revenue of $142,000
B. Customer satisfaction ratings on a 1-10 scale
C. Customer feedback describing a product as "difficult to use"
D. Website page views per day

✓ Correct Answer: C

Qualitative data describes qualities or characteristics in non-numerical form. Written customer feedback is a classic example — it cannot be directly averaged or summed.

Chapter 2

Ask Questions to Make Data-Driven Decisions

SMART questions, structured thinking, and stakeholder communication

Why the Right Questions Matter More Than the Right Answers

The most common mistake early analysts make is jumping straight to analysis without clearly defining what question they're answering. A well-defined question determines what data you need, how you clean it, and what your analysis should produce. Vague questions produce vague insights.

SMART Questions

The SMART Framework

Every good analytical question should be: Specific, Measurable, Action-oriented, Relevant, and Time-bound.

Letter	Meaning	Example
Specific	Clearly defined, not vague	"What was the average order value for customers in the Northeast in Q3?" vs. "How are sales?"
Measurable	Can be quantified or evaluated	"What percentage of users completed checkout?" vs. "Are users happy?"
Action-oriented	Leads to a decision or change	"Which marketing channel drives the highest ROI?" vs. "What is our marketing data?"
Relevant	Directly addresses the business problem	Tied to the decision you're helping make
Time-bound	Has a defined time period	"In the last 90 days" or "during Q4 2024"

Structured Thinking

Structured thinking means approaching problems systematically: define the problem, identify what you know and don't know, organize available information, and identify the gaps that need to be filled. This is how experienced analysts approach new projects rather than diving directly into data.

A useful tool for structured thinking is a **scope of work (SOW)** — a document that defines what analysis will be done, what data will be used, what deliverables are expected, and what the timeline is. Even for informal projects, writing one forces clarity.

Working with Stakeholders

Stakeholders are anyone with an interest in the outcome of your analysis — the person who requested it, the team that will act on it, leadership who will make decisions based on it. Understanding who your stakeholders are and what they actually need (which isn't always what they ask for) is one of the most important analyst skills.

Key Stakeholder Skills

- Ask clarifying questions before starting — it's faster than redoing work after the fact.
- Manage expectations about timelines and what the data can and cannot tell you.
- Communicate in plain language — not every stakeholder understands SQL or statistical significance.
- Proactively share roadblocks — don't go dark when you hit a problem.
- Present findings at the right level of detail for your audience.

Spreadsheet Fundamentals

Course 2 introduces spreadsheets as an analyst tool. The fundamentals you'll need:

Navigation and Formatting

Freeze rows and columns to keep headers visible while scrolling. Use named ranges to make formulas more readable. Apply consistent number and date formatting across your data to avoid cleaning headaches later.

Essential Formulas Introduced in Course 2

Formula	What It Does
=SUM(range)	Adds all values in a range
=AVERAGE(range)	Calculates the mean of a range
=COUNT(range)	Counts cells with numbers
=COUNTA(range)	Counts all non-empty cells
=MAX(range) / =MIN(range)	Returns the highest or lowest value
=IF(condition, value_if_true, value_if_false)	Returns different values based on a condition

Chapter 2 Practice Questions

Question 1

Which of these is the BEST example of a SMART analytical question?

A. How are our customers doing?

B. What was the average customer satisfaction score for online orders in Q4 2024, and how did it compare to Q3?

C. Why do customers leave?

D. Can you look into our sales data?

✓ Correct Answer: B

This question is Specific (online orders, Q4 2024), Measurable (average satisfaction score), Action-oriented (comparison that could drive decisions), Relevant (customer satisfaction), and Time-bound (Q4 2024 vs Q3).

Question 2

What is the primary purpose of a scope of work (SOW) in a data project?

A. To document the SQL queries used in the analysis

B. To define what the analysis will cover, what data will be used, and what deliverables are expected

C. To store raw data before it is cleaned

D. To create the final visualization for stakeholders

✓ Correct Answer: B

A scope of work defines the boundaries and expectations of a project upfront, preventing scope creep and ensuring analyst and stakeholder are aligned before work begins.

Chapter 3

Prepare Data for Exploration

Data types, databases, SQL introduction, and evaluating credibility

Understanding Data Sources

Before you analyze data, you need to understand where it came from and whether you can trust it. Data comes from two primary source categories:

First-party data is data your organization collected directly — customer purchase history, survey responses, app usage logs. It's the most reliable because you control the collection methodology.

Second-party data is data collected by another organization and shared directly with you — a partner's customer data shared under a data agreement. Generally trustworthy, but you have less control over collection quality.

Third-party data is collected by an outside organization and sold or shared publicly — census data, market research reports, social media aggregators. Useful for context but requires careful evaluation of methodology.

Data Bias and Credibility

Key Principle

Bad data leads to bad decisions even with perfect analysis. Evaluating data credibility before you start is not optional — it's the job.

The certificate uses the **ROCCC framework** to evaluate data credibility:

ROCCC Letter	Question to Ask
Reliable	Is the data accurate, complete, and unbiased?
Original	Is this from the primary source, or has it been filtered or aggregated by others?
Comprehensive	Does it contain all the information needed to answer the question?

ROCCC Letter	Question to Ask
Current	Is it recent enough to be relevant to your analysis?
Cited	Is the source clearly documented and credible?

Common Types of Data Bias

- Sampling bias — the sample doesn't represent the full population (e.g., surveying only online customers when you also have in-store customers).
- Observer bias — the presence of a researcher influences responses (e.g., in-person interviews vs. anonymous surveys).
- Confirmation bias — only looking for data that supports a pre-existing conclusion.
- Availability bias — over-relying on data that's easy to access rather than the most appropriate data.

Introduction to Databases and SQL

Most organizational data lives in relational databases — structured systems that store data in tables with rows and columns. SQL (Structured Query Language) is the standard language for retrieving and manipulating that data.

How a Relational Database Works

A relational database contains multiple tables. Tables are related to each other through shared columns called **keys**. A **primary key** uniquely identifies each row in a table. A **foreign key** in one table references the primary key of another, creating the relationship between them.

Your First SQL Queries

```
-- Select all columns from a table SELECT * FROM sales_data;
```

```
-- Select specific columns SELECT customer_id, order_total, order_date FROM sales_data;
```

```
-- Filter rows with WHERE SELECT * FROM sales_data WHERE order_total > 100;
```

```
-- Multiple conditions with AND SELECT * FROM sales_data WHERE order_total > 100 AND region = 'North';
```

Data Formats and File Types

Format	What It Is
.csv	Comma-separated values. The most common format for sharing tabular data.
.xlsx	Microsoft Excel format. Supports formulas, formatting, and multiple sheets.
.json	JavaScript Object Notation. Common in web APIs and app data exports.
.sql	A file containing SQL queries or database export commands.
.parquet	Columnar storage format used in big data environments (Spark, BigQuery).

Chapter 3 Practice Questions

Question 1

Which data source type is generally considered the most reliable?

A. Third-party data purchased from a market research firm
B. First-party data collected directly by your organization
C. Second-party data shared by a partner
D. Public government data

✓ Correct Answer: B

First-party data is collected under conditions you control, making it the most reliable. You understand the methodology, the collection period, and any limitations.

Question 2

What does the "R" in the ROCCC framework stand for?

A. Relevant
B. Reliable
C. Recent
D. Reviewed

✓ Correct Answer: B

The ROCCC framework evaluates: Reliable, Original, Comprehensive, Current, and Cited.

Question 3

In a relational database, what is a primary key?

A. The most important column in a table

B. A unique identifier for each row in a table

C. The first column of every table

D. A column shared between two tables

✓ Correct Answer: B

A primary key uniquely identifies each row in a table. No two rows can have the same primary key value, and it cannot be NULL.

Chapter 4

Process Data from Dirty to Clean

Data cleaning in spreadsheets and SQL, validation, and documentation

What Is Dirty Data?

Dirty data is any data that is inaccurate, incomplete, inconsistent, duplicate, or irrelevant. It's the rule, not the exception — real-world data is almost never clean when you receive it. Estimates from industry practitioners suggest analysts spend 60–80% of their time cleaning data rather than analyzing it.

Types of Dirty Data

Type	Description	Example
Duplicate data	Same record appears more than once	Customer listed twice after a system migration
Outdated data	Information that is no longer accurate	Old phone numbers or addresses
Incomplete data	Missing values in required fields	NULL email addresses in a customer table
Incorrect data	Values that are wrong	Age field containing 500, negative sales values
Inconsistent data	Same information formatted differently	'New York', 'new york', 'NY' in the same column
Irrelevant data	Data that doesn't serve the analysis	Columns from a system export you don't need

Data Cleaning in Spreadsheets

Remove Duplicates

In Google Sheets: Data → Data Cleanup → Remove Duplicates. In Excel: Data → Remove Duplicates. Always make a copy of your data before removing anything — you can't undo a bulk deletion after saving.

Essential Cleaning Formulas

Formula	Purpose
=TRIM(text)	Removes extra spaces from the beginning, end, and middle of text
=CLEAN(text)	Removes non-printable characters from text
=UPPER(text) / =LOWER(text)	Converts text to all uppercase or all lowercase
=PROPER(text)	Capitalizes the first letter of each word
=LEN(text)	Returns the number of characters in a string — useful for spotting unexpected values
=IFERROR(formula, value)	Returns a specified value instead of an error if a formula fails
=SUBSTITUTE(text, old, new)	Replaces specific characters or strings within a cell

Conditional Formatting for Spotting Issues

Use conditional formatting (Format → Conditional Formatting) to visually highlight outliers, blanks, or values that don't match expected patterns. For example, highlight any cells in an age column where the value is over 120 or below 0 — these are clearly errors.

Data Cleaning in SQL

Finding and Handling NULL Values

```
-- Find rows with NULL email SELECT * FROM customers WHERE email IS NULL; -- Count NULLs in a column SELECT COUNT(*) - COUNT(email) AS null_count FROM customers;
```

Finding Duplicates

```
-- Identify duplicate customer IDs SELECT customer_id, COUNT(*) as occurrences FROM orders GROUP BY customer_id HAVING COUNT(*) > 1;
```

Cleaning Text in SQL

```
-- Remove extra whitespace SELECT TRIM(customer_name) AS clean_name FROM customers; -- Standardize to lowercase SELECT LOWER(email) AS normalized_email FROM customers; -- Replace a value SELECT REPLACE(phone, '-', '') AS clean_phone FROM customers;
```

Critical Best Practice

Always document your cleaning process. Write down what you changed, why you changed it, and what the original data looked like. Your future self — and your colleagues — will thank you.

Chapter 4 Practice Questions

Question 1

A column in your dataset contains both "New York" and "new york" as values for the same city. What type of dirty data is this?

A. Duplicate data
B. Incomplete data
C. Inconsistent data
D. Outdated data

✓ Correct Answer: C

Inconsistent data occurs when the same information is represented differently across a dataset. Standardizing text case is one of the most common data cleaning tasks.

Question 2

Which SQL function removes extra whitespace from a string?

A. CLEAN()
B. STRIP()
C. TRIM()
D. REMOVE()

✓ Correct Answer: C

TRIM() removes leading and trailing whitespace from a string in SQL. In spreadsheets, TRIM() also removes extra spaces between words.

Question 3

What is the purpose of IFERROR() in a spreadsheet?

A. It prevents formulas from running if there is an error in the data
B. It returns a specified value instead of displaying an error if a formula fails

C. It identifies which cells contain errors

D. It automatically corrects errors in a dataset

✓ Correct Answer: B

IFERROR() is used to handle errors gracefully in formulas — for example, =IFERROR(VLOOKUP(A1, B:C, 2, FALSE), "Not Found") returns "Not Found" instead of #N/A when no match is found.

Chapter 5

Analyze Data to Answer Questions

SQL aggregations, JOINs, pivot tables, and VLOOKUP

The Core Analytical Mindset

Analysis is the process of finding patterns, connections, and insights in your data. Good analysis starts with a clear question and works backward to identify exactly what computation or query will answer it — not the other way around.

SQL: Aggregations and GROUP BY

Aggregation is the process of summarizing multiple rows of data into a single result. It's one of the most commonly tested skills in the GDA certificate.

Aggregate Functions

Function	What It Does	Example Usage
COUNT(*)	Counts all rows in a group	COUNT(*) AS total_orders
COUNT(column)	Counts non-NULL values in a column	COUNT(email) AS emails_on_file
SUM(column)	Adds all values in a column	SUM(order_total) AS revenue
AVG(column)	Calculates the mean	AVG(order_total) AS avg_order
MAX(column)	Returns the highest value	MAX(order_date) AS latest_order
MIN(column)	Returns the lowest value	MIN(price) AS cheapest_item

GROUP BY and HAVING

```
-- Count orders per region SELECT region, COUNT(*) AS order_count FROM sales
GROUP BY region ORDER BY order_count DESC; -- HAVING: filter groups (not rows)
```

```
SELECT region, AVG(order_total) AS avg_order FROM sales GROUP BY region HAVING AVG(order_total) > 75;
```

WHERE vs. HAVING — The Most Common Exam Mistake

WHERE filters individual rows BEFORE grouping. HAVING filters groups AFTER GROUP BY. You cannot use WHERE to filter aggregate results.

SQL JOINs

JOINs combine rows from two or more tables based on a related column. Understanding what each JOIN returns is critical for the certificate.

JOIN Type	Returns	When to Use
INNER JOIN	Returns only rows with matching values in BOTH tables	The most common join type
LEFT JOIN	Returns ALL rows from the left table, matching rows from the right (NULLs where no match)	Use when you want to keep all records from one table
RIGHT JOIN	Returns ALL rows from the right table, matching rows from the left	Less common; can usually be rewritten as a LEFT JOIN
FULL OUTER JOIN	Returns all rows from both tables, NULLs where no match	Use to find unmatched records in either table

```
-- INNER JOIN example SELECT o.order_id, c.customer_name, o.order_total FROM orders o INNER JOIN customers c ON o.customer_id = c.customer_id; -- LEFT JOIN: keep all customers even with no orders SELECT c.customer_name, COUNT(o.order_id) AS order_count FROM customers c LEFT JOIN orders o ON c.customer_id = o.customer_id GROUP BY c.customer_name;
```

Spreadsheet Analysis: Pivot Tables

Pivot tables are one of the most powerful and most-tested features in the certificate. They allow you to summarize, reorganize, and aggregate large datasets without writing formulas.

How to Create a Pivot Table

- Select your data range (include headers)
- In Google Sheets: Insert → Pivot Table. In Excel: Insert → PivotTable
- Choose where to place it (new sheet recommended)
- Drag fields to Rows, Columns, and Values sections
- Change the aggregation in Values (Sum, Count, Average, etc.)

VLOOKUP

VLOOKUP (Vertical Lookup) searches the first column of a range for a match and returns a value from a specified column in the same row.

```
=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup]) -- Example: Find price for product ID in A2 =VLOOKUP(A2, Products!A:C, 3, FALSE) -- Parameters: -- A2 = value to search for -- Products!A:C = range to search in -- 3 = return the value from the 3rd column -- FALSE = exact match (always use FALSE for most cases)
```

Chapter 5 Practice Questions

Question 1

What is the difference between WHERE and HAVING in SQL?

A. WHERE filters columns; HAVING filters rows
B. WHERE filters rows before grouping; HAVING filters groups after GROUP BY
C. They are interchangeable
D. WHERE is used with JOINs; HAVING is used with subqueries

✓ Correct Answer: B

This is one of the most tested SQL concepts in the certificate. WHERE filters individual rows before any grouping occurs. HAVING filters the results of GROUP BY aggregations.

Question 2

Which SQL JOIN returns ALL rows from both tables, with NULLs where there is no match?

A. INNER JOIN
B. LEFT JOIN
C. RIGHT JOIN
D. FULL OUTER JOIN

✓ Correct Answer: D

FULL OUTER JOIN returns every row from both tables, filling in NULLs where there is no matching record in the other table. This is useful for finding records that exist in one table but not the other.

Question 3

In VLOOKUP, what does setting the range_lookup parameter to FALSE do?

A. Returns an approximate match

B. Returns an exact match only

C. Searches the entire spreadsheet

D. Sorts the results alphabetically

✓ Correct Answer: B

FALSE (or 0) in the range_lookup parameter forces VLOOKUP to find an exact match. If no exact match exists, it returns #N/A. This is the correct setting for most analytical use cases.

Chapter 6

Share Data Through the Art of Visualization

Tableau, dashboard design, and data storytelling

Why Visualization Matters

Your analysis is only as valuable as your ability to communicate it. A perfect SQL query that produces an insight no one understands is worthless. Visualization bridges the gap between what the data shows and what stakeholders can act on.

Choosing the Right Chart Type

The Golden Rule

Choose the chart type that makes your key insight immediately obvious to someone who has never seen the data before. If it requires explanation, redesign it.

Goal	Best Chart Type	Example Use Case
Compare categories	Bar / Column chart	Sales by region, product performance
Show change over time	Line chart	Monthly revenue, weekly signups
Show distribution	Histogram	Age distribution, order value spread
Show part-to-whole	Pie / Donut chart	Market share (max 5–6 segments)
Show correlation	Scatter plot	Price vs. sales volume
Compare multiple variables	Heatmap / Bubble chart	Performance across regions and products
Show ranking	Horizontal bar chart	Top 10 customers by revenue

Tableau Fundamentals

The certificate uses Tableau Public (the free desktop version). The key concepts you need to know:

Dimensions vs. Measures

Dimensions are categorical fields — Region, Product Name, Customer Segment. They define how data is segmented or sliced. **Measures** are numeric fields that can be aggregated — Revenue, Order Count, Profit. Tableau automatically categorizes fields, but you can override this.

Building Basic Charts in Tableau

- Connect to your data source (CSV, Excel, Google Sheets, or database)
- Drag a dimension to the Columns shelf and a measure to the Rows shelf (or vice versa)
- Tableau auto-suggests a chart type — use Show Me to change it
- Drag additional dimensions to Color, Size, or Label to add visual encoding
- Use Filters to narrow your view to specific segments or time periods

Building Dashboards

A dashboard combines multiple charts into a single view. Key dashboard design principles: include no more than 3–5 charts per dashboard, add a filter so users can explore the data themselves, use a consistent color scheme, and always include a title that states the insight — not just the data.

Data Storytelling

Data storytelling combines data, visuals, and narrative to communicate insights in a way that drives understanding and action. The three elements must work together:

Element	Role
Data	The evidence that supports your insight. Must be accurate and relevant.
Visuals	Charts and graphs that make the pattern immediately visible.
Narrative	The context and meaning — why this matters and what to do about it.

Presenting to Stakeholders

Lead with the insight, not the methodology. "Sales in the Northeast dropped 18% in Q3" is a better opening than "I ran a SQL query that joined the orders and regions tables." Structure your presentation: key finding → supporting evidence → recommendation → next steps.

Chapter 6 Practice Questions

Question 1

Which chart type is BEST for showing how a single numeric variable is distributed across a range?

A. Bar chart
B. Line chart
C. Histogram
D. Pie chart

✓ Correct Answer: C

Histograms display the frequency distribution of a continuous numeric variable by grouping values into bins. They show the shape of the distribution — whether it's normal, skewed, or bimodal.

Question 2

In Tableau, what is the difference between a dimension and a measure?

A. Dimensions are larger datasets; measures are smaller
B. Dimensions are categorical fields used to segment data; measures are numeric values that can be aggregated
C. Measures are calculated fields; dimensions are raw fields
D. They are interchangeable terms

✓ Correct Answer: B

Dimensions define how data is sliced and categorized (Region, Product Type). Measures are numeric values you aggregate (SUM of Revenue, COUNT of Orders). Understanding this distinction is fundamental to working in Tableau.

Question 3

What is the primary goal of data storytelling?

A. To demonstrate the analyst's technical proficiency
B. To display as much data as possible in a single view
C. To communicate insights clearly so stakeholders can make informed decisions
D. To create the most visually complex chart possible

✓ Correct Answer: C

Data storytelling is about driving understanding and action, not showcasing technical skill. The best visualizations are ones that make an insight immediately clear to a non-technical audience.

Chapter 7

Data Analysis with R Programming

R basics, tidyverse, ggplot2, and R Markdown

Why R?

R is a programming language specifically designed for statistical analysis and data visualization. While Python is more popular in data science and machine learning, R remains widely used in academic research, healthcare analytics, and organizations with deep statistical analysis needs. The GDA certificate introduces R at a practical, beginner level.

Getting Started with RStudio

RStudio is the integrated development environment (IDE) for R. Think of it as the tool you use to write and run R code. Download RStudio Desktop (free) from posit.co. The interface has four main panes:

- Console — where R runs your code and shows output
- Script Editor — where you write and save your R scripts
- Environment — shows all objects currently loaded in memory
- Files / Plots / Packages / Help — utility panes for navigation and documentation

R Basics

Variables and Data Types

```
# Assign a value to a variable my_name <- "Alex" my_age <- 28 is_enrolled <- TRUE # Check the type of a variable class(my_age) # Returns "numeric" class(my_name) # Returns "character"
```

Loading Data

```
# Load a CSV file my_data <- read.csv("sales_data.csv") # Preview the data head(my_data) # First 6 rows glimpse(my_data) # Column names, types, and
```

```
sample values str(my_data) # Structure overview
```

The Tidyverse

The tidyverse is a collection of R packages designed for data analysis. Install it once with install.packages("tidyverse"), then load it with library(tidyverse). The most important packages for the GDA certificate:

Package	Purpose
dplyr	Data manipulation — filter, select, mutate, group_by, summarize
ggplot2	Data visualization — build charts layer by layer
tidyr	Data reshaping — pivot data between wide and long formats
readr	Reading data files — faster and cleaner than base R

Essential dplyr Functions

```
library(tidyverse) # Filter rows filtered_data <- my_data %>% filter(region == "North", sales > 1000) # Select specific columns selected_data <- my_data %>% select(customer_id, sales, region) # Create a new column mutated_data <- my_data %>% mutate(profit_margin = (revenue - cost) / revenue) # Summarize by group summary_data <- my_data %>% group_by(region) %>% summarize(total_sales = sum(sales), avg_sales = mean(sales))
```

ggplot2 Visualizations

ggplot2 builds charts in layers using the grammar of graphics. Every chart starts with ggplot() defining the data and aesthetic mappings, followed by a geom_ function defining the chart type.

```
# Basic bar chart ggplot(data = sales_summary, aes(x = region, y = total_sales)) + geom_bar(stat = "identity", fill = "#00c2a8") + labs(title = "Total Sales by Region", x = "Region", y = "Total Sales ($)") + theme_minimal() # Scatter plot ggplot(data = my_data, aes(x = ad_spend, y = revenue)) + geom_point(color = "#0b1f3a", alpha = 0.6) + geom_smooth(method = "lm", color = "#f5c542") + labs(title = "Ad Spend vs. Revenue")
```

R Markdown

R Markdown lets you combine your R code, output, and written narrative in a single document that can be exported as HTML, PDF, or Word. It's the standard way to present R analysis professionally. For the GDA certificate capstone, R Markdown is an excellent format for documenting your case study.

Chapter 7 Practice Questions

Question 1

What does the pipe operator (%>%) do in R's tidyverse?

A. It divides one number by another
B. It passes the output of one function as the input to the next function
C. It creates a new data frame
D. It loads a package into the environment

✓ Correct Answer: B

The pipe operator %>% (from the magrittr/dplyr package) passes the result of the left-hand side as the first argument to the right-hand side. It makes code more readable by chaining operations in a logical left-to-right flow.

Question 2

Which ggplot2 function creates a scatter plot?

A. geom_bar()
B. geom_scatter()
C. geom_point()
D. geom_line()

✓ Correct Answer: C

geom_point() creates a scatter plot in ggplot2. Each row in your data becomes a point on the chart, with x and y positions defined by your aes() mappings.

Chapter 8

Career Prep & Capstone

Capstone project, portfolio, resume, LinkedIn, and interview prep

The Capstone Project

The capstone is the most important deliverable in the GDA certificate. It's the project that anchors your portfolio, demonstrates your skills to employers, and gives you something concrete to discuss in interviews. Treat it with the same seriousness you would treat a real work assignment.

Choosing Your Case Study

The certificate offers two pre-built case studies (Cyclistic bike share and Bellabeat wellness products). You can also design your own using public datasets. The pre-built cases are fine — but using a dataset related to your target industry makes your capstone more relevant in job applications.

Good sources for custom datasets: Kaggle.com, Google Dataset Search (datasets.google.com), data.gov, Our World in Data, and industry-specific government portals.

Structuring Your Capstone

Use the GDA data analysis framework explicitly in your write-up:

Phase	What to Include
Ask	State the business problem and the specific questions your analysis answers
Prepare	Describe your data source, format, time period, and any limitations
Process	Document what cleaning you did and why. Include before/after examples
Analyze	Show your key queries, calculations, or R code. Explain your reasoning
Share	Include at least 3 visualizations. Use titles that state the insight
Act	Provide 3–5 specific, actionable recommendations based on your findings

Building Your Portfolio

A portfolio is your proof of work — it shows employers you can do the job, not just pass tests about it. Your capstone is the foundation, but aim for 3–5 projects total.

What to Include in Each Project

- A clear problem statement — what business question were you answering?
- Data source description and any limitations
- Key cleaning steps and decisions
- Analysis code (SQL queries or R scripts) with comments
- At least 2–3 visualizations with insight-focused titles
- Findings and recommendations in plain language

Where to Host Your Portfolio

- GitHub — the standard for technical portfolios. Create a repo for each project with a detailed README.
- Tableau Public — for visualization-focused projects. Dashboards are publicly viewable.
- Personal website — optional but differentiating. GitHub Pages is free.
- Google Sites or Notion — simple options if you prefer no-code hosting.

Resume Guide for GDA Certificate Holders

How to List the Certificate

Create a Certifications section near the top of your resume (after your summary):

Certificate Format

Google Data Analytics Certificate, Google / Coursera, [Year Completed]

Skills Section

List the tools and skills you can discuss confidently in an interview: SQL (BigQuery, MySQL), Spreadsheets (Google Sheets, Excel), Data Visualization (Tableau), R Programming, Data Cleaning, Pivot Tables, VLOOKUP, Statistical Analysis, Dashboard Design.

Writing Strong Bullet Points

Quantify impact wherever possible. Weak: "Analyzed sales data in SQL." Strong: "Analyzed 50,000+ transaction records in BigQuery to identify $180K revenue opportunity in underserved customer segment; findings adopted by marketing team for Q4 campaign targeting."

LinkedIn Optimization

Your Headline

Don't waste your headline with just your job title. Use it to signal what you do and what you're targeting: "Data Analyst | SQL · Tableau · Google Data Analytics Certificate | Open to Opportunities"

Featured Section

Use the Featured section to link directly to your GitHub portfolio or Tableau Public profile. This is prime real estate that most candidates ignore — a link to a real project here is more persuasive than three more bullet points.

Interview Preparation

Technical Questions to Prepare For

Topic	Sample Question
SQL	Write a query to find the top 5 customers by total spend. How would you handle duplicate records?
Spreadsheets	How would you use VLOOKUP to combine data from two sheets? Walk me through a pivot table.
Data Cleaning	How do you handle missing values? How do you document your cleaning process?
Visualization	How do you decide which chart type to use? Walk me through a dashboard you built.
Process	Walk me through how you would approach a new analysis request from a stakeholder.

The STAR Method for Behavioral Questions

Behavioral questions ("Tell me about a time when...") are answered most effectively using the STAR format: **Situation** (context), **Task** (what you needed to do), **Action** (what you specifically did), **Result** (the outcome, ideally quantified). Prepare 3–5 STAR stories using your capstone and portfolio projects.

Chapter 8 Practice Questions

Question 1

Which of the following is the BEST headline for a data analyst job seeker on LinkedIn?

A. Recent Graduate

B. Data Analyst | SQL · Tableau · Google Data Analytics Certificate | Open to Opportunities

C. Looking for a job in data

D. Student at Coursera

✓ Correct Answer: B

A strong LinkedIn headline includes your target role, key skills, relevant credentials, and availability. It is keyword-rich (SQL, Tableau, Data Analyst are searchable terms) and immediately communicates your value proposition.

Question 2

When presenting analysis results to non-technical stakeholders, what should you lead with?

A. The SQL queries you used

B. The key insight or finding

C. A detailed description of the dataset

D. The tools and methods you employed

✓ Correct Answer: B

Non-technical stakeholders care about the business implication, not the methodology. Lead with the insight ("Sales dropped 18% in Q3 driven by the Northeast region"), then provide supporting evidence and methodology if asked.

BONUS: SQL Quick Reference Sheet

Core SQL Commands

Command	Example	What It Does
SELECT	SELECT col1, col2 FROM table	Choose which columns to return
WHERE	WHERE price > 100	Filter rows by condition
ORDER BY	ORDER BY date DESC	Sort results (ASC or DESC)
LIMIT	LIMIT 10	Return only N rows
DISTINCT	SELECT DISTINCT city	Remove duplicate values
GROUP BY	GROUP BY region	Group rows for aggregation
HAVING	HAVING COUNT(*) > 5	Filter groups after GROUP BY
JOIN	JOIN table2 ON t1.id = t2.id	Combine rows from multiple tables
AS	COUNT(*) AS order_count	Rename a column in results
IS NULL	WHERE email IS NULL	Find missing values
BETWEEN	WHERE age BETWEEN 18 AND 35	Filter within a range
LIKE	WHERE name LIKE 'A%'	Pattern matching (% = wildcard)
IN	WHERE region IN ('North','South')	Match any value in a list
CASE	CASE WHEN x > 0 THEN 'Pos' ELSE 'Neg' END	Conditional logic

SQL Query Order of Operations

SQL clauses must appear in this order in your query, and they are also processed in a specific internal order:

1. Write Order	SELECT → FROM → JOIN → WHERE → GROUP BY → HAVING → ORDER BY → LIMIT

2. Processing Order	FROM → JOIN → WHERE → GROUP BY → HAVING → SELECT → ORDER BY → LIMIT

BONUS: Spreadsheet Formula Cheat Sheet

Lookup & Reference

Formula	What It Does
=VLOOKUP(val, range, col, FALSE)	Find a value in the first column and return a value from another column in the same row
=HLOOKUP(val, range, row, FALSE)	Same as VLOOKUP but searches horizontally
=INDEX(range, row, col)	Returns the value at a specific row/column intersection
=MATCH(val, range, 0)	Returns the position of a value within a range
=INDEX(range, MATCH(val, col, 0))	Powerful alternative to VLOOKUP — works with any column

Logical

Formula	What It Does
=IF(condition, true, false)	Returns one value if true, another if false
=IFS(cond1, val1, cond2, val2)	Multiple conditions without nested IFs
=AND(cond1, cond2)	Returns TRUE if ALL conditions are met
=OR(cond1, cond2)	Returns TRUE if ANY condition is met
=IFERROR(formula, value)	Returns value instead of error if formula fails

Counting & Summing

Formula	What It Does
=COUNT(range)	Counts cells containing numbers
=COUNTA(range)	Counts all non-empty cells
=COUNTIF(range, criteria)	Counts cells meeting a condition
=COUNTIFS(r1, c1, r2, c2)	Counts cells meeting multiple conditions

Formula	What It Does
=SUMIF(range, criteria, sum_range)	Sums cells that meet a condition
=SUMIFS(sum_r, r1, c1, r2, c2)	Sums cells meeting multiple conditions
=AVERAGEIF(range, criteria, avg_r)	Averages cells meeting a condition

Text

Formula	What It Does
=LEFT(text, n)	Returns first n characters
=RIGHT(text, n)	Returns last n characters
=MID(text, start, n)	Returns n characters starting at position start
=LEN(text)	Returns number of characters in a string
=TRIM(text)	Removes extra spaces
=UPPER/LOWER/PROPER(text)	Changes text case
=CONCATENATE(t1, t2) or t1&t2;	Joins text strings together
=SUBSTITUTE(text, old, new)	Replaces text within a string